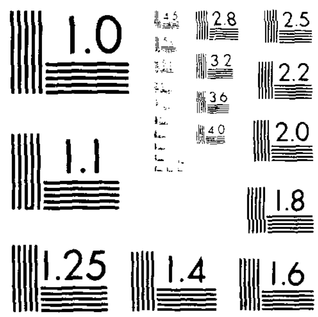
AD-A085 379 MINNESOTA UNIV ST PAUL DEPT OF APPLIED STATISTICS F/G 12/1
FISHER'S CONTRIBUTIONS TO THE ANALYSIS OF CATEGORICAL DATA.(U)
1980 S E FIENBERG, D V HINKLEY N00014-78-C-0600
UNCLASSIFIED NL

1 OF 1
AD A085379

END
DATE
FILMED
7-80
DTIC



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A

Appearing in R.A. Fisher: An Appreciation, edited by S.E. Fienberg and D.V. Hinkley (New York: Springer-Verlag), 1980.

Contract N00014-78-C-0600

Minnesota Univ., St. Paul

FISHER'S CONTRIBUTIONS TO THE ANALYSIS OF CATEGORICAL DATA

Stephen E. Fienberg

1. Introduction

Those who have had only the briefest of introductions to the analysis of categorical data will have learned of Fisher's contributions such as his exact test. But Fisher's work in this area covered a variety of topics, including fundamental papers on the distribution of the chi-square statistic which brought him into a major confrontation with Karl Pearson.

For years Pearson and his students had been applying the chi-square goodness-of-fit test incorrectly to a wide variety of problems. In a series of five papers, beginning in 1922 and continuing through 1928, Fisher wrote about the χ^2 method, correcting Pearson's errors. Specifically he explained how the appropriate degrees of freedom for χ^2 were to be calculated, and why the use of maximum likelihood and other efficient methods of estimation were intimately related to the χ^2 ideas. Of course, these papers were written during the period when Fisher was publishing his landmark papers on statistical estimation and the foundations of statistics [CP 42] (see the related discussion in the presentations by Hinkley in this volume), and thus the link to the more general theory was a natural one.

This presentation will concentrate on the five papers on χ^2, but will also discuss other related papers published by Fisher subsequently. It is interesting to note that the recent literature on loglinear models for categorical data problems can also be traced directly to Fisher, although he did not actually write about methods for multiway tables. In his pioneering paper on 2·2·2 tables, Bartlett (1935) attributes to Fisher the idea of using the equality of Yule's cross-product ratios as a model.

In the next section we begin by reviewing Pearson's work on χ^2.

DISCLAIMER NOTICE

THIS DOCUMENT IS BEST QUALITY PRACTICABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

2. Historical Background

To gain a perspective on Fisher's contributions to the theory of χ^2 and the analysis of categorical data, we need to begin with Pearson's (1900) paper. In that paper, Pearson looked at the problem of comparing a set of observed and expected frequencies through the use of the χ^2 statistic,

$$\chi^2 = \sum_1^{n'} \frac{(x-m)^2}{m} = \sum_1^{n'} \frac{e^2}{m}, \qquad (1)$$

where m are the expected values (known), x the observed values, n' the number of cells, and e = x-m which are subject to the constraint $\Sigma e = 0$ (i.e., $\Sigma x = \Sigma m$). Pearson assumed that $\underset{\sim}{x}' = (x_1,\ldots,x_{n'})$ has a multinomial distribution which is well-approximated by a multivariate normal. He then showed that X^2 is a quadratic form that has a χ^2 distribution with n' - 1 degrees of freedom. With this result in hand, Pearson went on to consider the case when the m's are not known a priori, but are fitted by a model using the data. His argument went roughly as follows.

Let m be the true expected cell value and m_s the corresponding sample-based estimate, such that $m = m_s + \mu$ where μ is considered to be small. Then, if we ignore terms of order $(\mu/m_s)^3$, we have the following approximation to $X^2 - X_s^2$, where $X_s^2 = \sum_1^{n'} (x-m_s)^2/m_s$:

$$X^2 - X_s^2 = -\sum_1^{n'} \frac{\mu(x^2-m_s^2)}{m_s^2} + \sum_1^{n'} \left(\frac{\mu}{m_s}\right)^2 \frac{x^2}{m_s}. \qquad (2)$$

Pearson claimed that this difference is negligible when a large sample is considered. Thus he argued that, in the case of a large sample, X_s^2 also has an approximate χ^2 distribution with n' - 1 degrees of freedom.

This result is obviously wrong. Greenwood and Yule (1915) gave an argument against it by noting that in the approach of comparing two proportions, the square of the statistic used behaves more like a χ^2 with 1 degree of freedom rather than 3. But it took Fisher to present a more carefully reasoned argument.

3. The 1922 Paper [CP 19]

Fisher in his first paper on χ^2 in 1922 gives the correct solution to the Pearson χ^2 problem with estimated expected frequencies. He begins by noti g t the χ^2 test as advocated by Pearson is used in "contingency tables in which th sum of the deviations in any row and column is necessarily zero". He then jum and immediately identifies the degrees of freedom in an r×c table as (r-1)(c-1 Fisher argues that "the values of e can be regarded as independent co-ordinate generalized space, lying in a subspace of dimension equal to the degrees of f due to the linear constraints". The discussion here is not really a "proof" and includes some circular reasoning -- Fisher's intuition, however, was most certainly correct. Fisher also shows that a normal statistic for testing p_1 in a 2-binomial problem, when squared, is identical to the χ^2 statistic. (Gr wood and Yule conjectured this result, but Fisher was the first to outline a proof.) This result gives some support to his solution of the general χ^2 prob with estimated parameters. Finally, he notes that for a 2×s table, the associa degrees of freedom is s-1. This yields the same result as the correction propo by Pearson to solve the problem of independence in a 2×s contingency table. Ye Pearson never really explained why the 2×s table should be treated differently f the standard r×c contingency table problem.

In his prefatory note for the Collected Papers, Fisher describes this paper as follows:

> This short paper, with all its juvenile inadequacies, yet did something to break the ice. Any reader who feels exasperated by its tentative and piecemeal character should remember that it had to find its way to publication past critics who, in the first place, could not believe that Pearson's work stood in need of correction, and who, if this had to be admitted, were sure that they themselves had corrected it.

4. The 1923 Paper [CP 31]

While the arguments in the 1922 paper may now seem clear and understandable to us today, they led to a predictable controversy in the 1920s. Thus Fisher elaborated his ideas in [CP 31], stressing the concept of degrees of freedom adjusted for the estimation of parameters. The paper begins with the following cross-classification of problems of interest:

	Random Sample	Selected Sample
True Population	A. No correction needed	B. Correction needed
Reconstructed Population	C. Correction needed	---------

Case A (Pearson's original problem) is the basic situation of multinomial sampling with known parameter values for which there was agreement among all concerned. Case B involves known linear restraints (e.g., fixed row totals in a 2×2 table), and again there was general agreement that an adjustment to the degrees of freedom was needed. Case C involves estimated parameters, and was the point of contention.

Fisher, after outlining the problem, goes on to describe a sampling experiment carried out by Yule (1922) for 2×2 tables, involving 350 observations for Case C. The distribution of the values of X^2 is included here in Table 1.

Table 1: The Distribution of the Values of X^2

	Number Expected, n' = 2	Number Observed	Number Expected, n' = 4
0–0.25	134.02 +	122	10.80 -
0.25–0.50	48.15 -	54	17.58 -
0.50–0.75	32.56 -	41	20.13 -
0.75–1.00	24.21 +	24	21.05 -
1–2	56.00 -	62	80.10 +
2–3	25.91 +	18	63.27 +
3–4	13.22 +	13	45.56 +
4–5	7.05 +	6	31.38 +
5–6	3.86 -	5	21.07 +
6—	5.01 +	5	39.06 +
	349.99	350	350

Fisher points out that "there can be no question that the expectation for n' = 4 completely fails while n' = 2 fits the observations well, and the correction is undoubtedly needed".

4. The 1924 Paper [CP 34]

The argument over the distribution of χ^2 was not settled by Fisher's 1922 and 1923 papers. Pearson (1922) denounced Fisher's claims, without referring specifically to him by name ("I trust my critic will pardon me for comparing him to Don

Quixote tilting at a windmill"). Fisher thus felt compelled to carry on the battle.

In his 1924 paper, Fisher shows with some care what was wrong with Pearson's original reasoning on the distribution of χ^2. He begins by describing three situations where we should not expect to achieve the usual asymptotic χ^2 distribution:

(A) the hypothesis tested is not in fact true,

(B) the method of estimation for the expected values is inconsistent,

(C) *the method of estimation employed is inefficient.*

Two properties of efficient estimates reviewed by Fisher in this context are worth mentioning here:

(1) The correlation between any two efficient estimates of the same parameter tends to one as the sample size tends to ∞.

(2) The correlation between an efficient and any other consistent estimate is $\sqrt{E}$ where E is the efficiency of the consistent estimate.

In this paper, Fisher also discusses minimizing the value of X^2 with respect to the parameter θ. He notes that the minimum is achieved when

$$\Sigma\left(\frac{x^2-m^2}{m^2}\right)\frac{\partial m}{\partial\theta} = 0. \tag{3}$$

By comparison the maximum likelihood estimate (MLE) satisfies the equations

$$\Sigma\left(\frac{x-m}{m}\frac{\partial m}{\partial\theta}\right) = 0. \tag{4}$$

Fisher claims that, for large samples, the factor $(x+m)/m$, by which the terms in (3) and (4) differ, tends in all cases to the value 2. Hence all methods involving any efficient statistic tend to minimize X^2. He then takes a new statistic, X'^2, equal to $\Sigma(x-m')^2/m'$ where m' is calculated using an efficient estimate, and finds the difference between X^2 and X'^2 as

$$X^2 - X'^2 = \Sigma\left[\frac{(x-m)^2}{m} - \frac{(x-m')^2}{m'}\right]$$

$$= \Sigma[x^2(\frac{1}{m} - \frac{1}{m'})]. \tag{5}$$

Also

$$\frac{1}{m} - \frac{1}{m'} = -\frac{1}{m'^2}\frac{\partial m'}{\partial\theta}\,\delta\theta + \left(\frac{2}{m'^3}\left(\frac{\partial m'}{\partial\theta}\right)^2 - \frac{1}{m'^2}\frac{\partial^2 m'}{\partial\theta^2}\right)\left(\frac{\delta\theta}{2}\right)^2, \tag{6}$$

where

$$\delta\theta = \theta - \theta' = O(n^{-\frac{1}{2}}). \tag{7}$$

But since X^2 has been made a minimum, we have

$$\Sigma\left(\frac{x^2}{m'^2}\frac{\partial m'}{\partial\theta}\right) = 0.$$

Thus expression (5) reduces to

$$X^2 - X'^2 = (\delta\theta)^2\,\Sigma\left[\frac{1}{m'}\left(\frac{\partial m'}{\partial\theta}\right)^2\right]$$

$$= \frac{(\delta\theta)^2}{\sigma^2(\theta')}, \tag{8}$$

and we get a reduction of X^2 when we estimate θ efficiently. Fisher also discussed the effects of estimating θ inefficiently, a research topic which has once again become fashionable in recent years (see the recent discussion in Fienberg, 1979).

5. The 1926 Paper [CP 49]

In his final paper attacking Karl Pearson's use of X^2, Fisher uses E.S. Pearson's experimental data on the distribution of binomial p, which for each of 12,448 different events contains the frequency of occurrence in two samples of 20 and 15, respectively. Fisher eliminates 780 cases of zero total occurrences, where X^2 is indeterminate, and for the remaining 11,668 cases computes the average X^2 for each value of total number of occurrences. The results are given in Table 2.

Table 2: Average X^2 for Each Value of Total Number of Occurrences

# of Successes	1	2	3	4	5	6	7	8	9
# of Tables	708	821	779	792	769	730	727	694	630
Total X^2	782.10	834.08	768.82	772.86	807.92	775.74	740.85	697.21	562.33
Average	1.0184	1.0159	0.9869	0.9758	1.0506	1.049[illegible]	1.0190	1.0046	0.8926

# of Successes	10	11	12	13	14	15	16	17	Total
# of Tables	643	670	682	668	616	568	524	578	11668
Total X^2	598.09	639.87	707.06	634.06	603.26	618.11	526.52	599.24	11668.17
Average	0.9302	0.9550	1.0368	0.9492	0.9793	1.0882	1.0048	1.0367	1.00[illegible]1

He notes that in every case the average value is "embarrassingly close" to 1, in no case is it near 3. To this paper, Pearson wrote no reply.

6. The 1928 and 1942 Papers [CP 62, 188]

His dispute with Pearson at least technically behind him, Fisher continued to write about categorical data problems and X^2, stressing the use of maximum likelihood for fitting the expected values.

In his 1928 paper, Fisher takes a genetic example with underlying cell probabilities $\frac{1}{4}(2+\theta, 1-\theta, 1-\theta, \theta)$ corresponding to observed frequencies (a_1, a_2, a_3, a_4). He notes that $x = a_1 + a_2 - 3(a_3 + a_4)$, $y = a_1 + a_3 - 3(a_2 + a_4)$ each have expectation zero for all values of θ, and that they may be identified with the two degrees of freedom available for testing goodness-of-fit of the model. The appropriate normal quadratic form in x and y, i.e., with their covariance matrix inverse as kernel, is

$$Q^2 = \frac{3}{8n(1-\theta)(1+2\theta)}\left(x^2 + y^2 - \frac{2}{3}(4\theta-1)xy\right). \tag{9}$$

He then compares Q^2 with the classical chi-square for given θ, namely

$$X^2 = \frac{4}{n}\left(\frac{a_1^2}{2+\theta} + \frac{a_2^2 + a_3^2}{1-\theta} + \frac{a_4^2}{\theta}\right) - n, \tag{10}$$

the difference being

$$X^2 - Q^2 = \left\{\frac{a_1}{2+\theta} - \frac{a_2+a_3}{1-\theta} + \frac{a_4}{\theta}\right\}^2 \Big/ \frac{(1+2\theta)n}{2\theta(1-\theta)(2+\theta)} .$$

Thus $X^2 - Q^2 > 0$ unless

$$\frac{a_1}{2+\theta} - \frac{a_2+a_3}{1-\theta} + \frac{a_4}{\theta} = 0. \qquad (11)$$

But expression (11) is exactly the likelihood equation. Thus the difference $X^2 - Q^2$ can be regarded as that part of the discrepancy between observation and hypothesis which is due to imperfect methods in the estimation of θ.

Fisher goes on to discuss maximum likelihood estimation (MLE) of expected frequencies in general with s cells and r parameters $\theta_1, \theta_2, \ldots, \theta_r$. He notes that the quadratic form analogous to expression (11) is made up of two parts, one of which is a quadratic form distributed in large samples as χ^2_{s-r-1}, and the other being due to errors of measurement, meaning inefficient estimation.

Likelihood estimation for categorical data problems continued to fascinate Fisher, and in 1942 [CP 188] he wrote a brief note on a χ^2 problem for which the likelihood solution comes out neatly. Let a, b, and c be three binomial variates with parameters p, p', and pp', and sample sizes A, B, and C, respectively. Then the value of X^2 with MLE's substituted for the expected values is

$$X^2 = \lambda^2 \left[\frac{A-a}{A(a+\lambda)} + \frac{B-b}{B(b+\lambda)} + \frac{C-c}{C(c-\lambda)}\right] , \qquad (12)$$

where λ is a root of

$$(A+\lambda)(B+\lambda)(C-\lambda) = (a+\lambda)(b+\lambda)(c-\lambda). \qquad (13)$$

Fisher extends this result to the case of s probabilities with s+1 binomial variates (the extra one corresponding to the combined event with probability the product of the s probabilities).

7. Confidence Limits for the Cross-Product Ratio [CP 291]

Even after his retirement to Australia, Fisher continued to write about methods for the 2×2 tables. In one of his last publications [CP 291], he briefly explored the use of the distribution of the exact test statistic (and χ^2 with

Yates' correction) to set limits on the population cross-product ratio. His example was as follows.

Let the observed table be

10	3
2	15

with expected frequencies

10-x	3+x
2+x	15-x

and cross-product ratio

$$\text{cpr} = \frac{(10-x)(15-x)}{(3+x)(2+x)} .$$

Now

$$X_c^2 = (x-\tfrac{1}{2})^2 \left(\frac{1}{10-x} + \frac{1}{3-x} + \frac{1}{2-x} + \frac{1}{15-x}\right).$$

Next we pick x to make $X^2 = 3.841$ (the 95th percentile), yielding $x = 3.0491$ and cpr = 2.720. The latter, Fisher argues, is an "upper limit" for the true cpr.

8. The Exact Test and χ^2

Fisher introduced his exact test for 2×2 tables with the now classic example of the lady tasting tea in The Design of Experiments [DOE, 1935] (see the discussion of this example in the lecture by Holschuh in this volume). He advocated the use of the exact test in subsequent issues of Statistical Methods for Research Workers [SMRW, 1925] (Section 21.02), referring to the use of χ^2 as an approximation. The χ^2 test for 2×2 tables with the correction for continuity introduced by Yates (used in Section 7 above) was an attempt to get tail probability values that conformed more closely to those of the exact test than did those from the uncorrected χ^2 statistic.

In a 1941 Science article [CP 183], in response to a paper by E.B. Wilson, Fisher tried to clarify why he believed that the exact test should be used in 2-binomial experiments when the sample sizes are small. The discussion in this

paper is obscure at best, and to the present day few have correctly described Fisher's position in a coherent fashion. Indeed his position on this issue (as on others) seems to have changed over time. Berkson (1978) and Kempthorne (1979) have continued the debate over the appropriateness of the exact test, and I fear that we will continue to see papers on this topic in the future.

My current judgment is that Fisher and others consistently overstated the dangers of using X^2 in small samples as if it really was distributed as a χ^2 variate (e.g., see the small sample studies of Larntz, 1978).

References

Bartlett, M.S. (1935). "Contingency Table Interactions," Supplement to the Journal of the Royal Statistical Society, 2, 248-252.

Berkson, J. (1978). "In Dispraise of the Exact Test," Journal of Statistical Planning and Inference, 2, 27-42.

Fienberg, S.E. (1979). "The Use of Chi-Squared Statistics for Categorical Data Problems," Journal of the Royal Statistical Society, Series B, 41, 54-64.

Greenwood, M. and G.U. Yule (1915). "The Statistics of Antityphoid and Anti-cholera Inoculations, and the Interpretation of Such Statistics in General," Proceedings of the Royal Society of Medicine, Section of Epidemiology and State Medicine, viii, 113.

Kempthorne, O. (1979). "In Dispraise of the Exact Test: Reactions," Journal of Statistical Planning and Inference, 3, 199-213.

Larntz, K. (1978). "Small-Sample Comparisons of Exact Levels for Chi-Squared Goodness-of-Fit Statistics," Journal of the American Statistical Association, 73, 253-263.

Pearson, K. (1900). "On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling," The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1, 157.

Pearson, K. (1922). "Further Note on the χ^2 Test of Goodness of Fit," Biometrika, 14, 418.

Yule, G.U. (1922). "On the Application of the χ^2 Method to Association and Contingency Tables with Experimental Illustration," Journal of the Royal Statistical Society, 5, 95-104.

This research was supported by Office of Naval Research Contract N00014-78-C-0600 to the University of Minnesota, St. Paul